Title
Abuse

Author
Roberts, Sarah

Publication Date
2020

Peer reviewed

Abuse

Sarah T. Roberts

Introduction: Content Moderation and Archives as Practice and Metaphor

Content moderation, or the adjudication of online user-generated content (UGC), is as much a practice of what is not seen as what is. Particularly when industrialized and done at scale, it directly impacts the landscape and ecology of social media platforms by being a bidirectional gatekeeping mechanism for both what is allowed to stay up as well as what is deleted. The latter can rarely ever be perceived or apprehended in a meaningful way, leading to negative consequences. In a study on user perceptions of online content moderation, Sarah Myers West (2018) found that such lack of transparency leads users to 'develop "folk theories" about how platforms work: in the absence of authoritative explanations, they strive to make sense of content moderation processes by drawing connections between related phenomena, developing non-authoritative conceptions of why and how their content was removed'.

Yet there is great significance that can be derived from the corpus of what might, in an unanalysed state, be considered piecemeal and individual digital errata/detritus (Roberts 2018). For this volume dedicated to the uncertain, unruly, disobedient archive, I probe the erasure en masse of

ostensibly abusive, problematic, dangerous and disturbing material from the landscape of the mainstream social Internet, the mechanisms that encourage its attempted circulation in the first place and its subsequent aggregation, under computational lock and key, in a digital repository archive—records captured (in this case, digital imagery and video) in a one-way relationship in which the material is intended to get in and never get out.

In order to apprehend the meaning of both the capturing process and the resulting archive, I connect the discussion to theoretical breakthroughs from the field of critical archival studies that have pushed back on conceptions of archival neutrality—on the field's homogeneity—that have changed the focus to community-oriented archives, and that have imagined liberatory, social justice-oriented and human rights frameworks for archives (Caswell 2014b; Punzalan and Caswell 2015; Sutherland 2017; Wood et al. 2014).

Such insights often unfairly suffer from a disciplinary cloistering lessening their key impact or uptake in discussions of, on the one hand, mainstream (and self-styled apolitical or neutral) archival practice, and on the other, sociological, anthropological and humanities takes on 'the archive' that often engage relatively little with archival theory and practice behind the object of study: the monolithic, abstract and often titular 'archive'. Yet this dialogue within critical archival studies has much to offer to nature and import of these corpora *in toto*—in theory and in practice. Here I am

indebted to that field's task of 'trying to read [the archive's] narratives of power and knowledge' (Ketelaar 2001, 132) in order to make sense of the case of one particular archive as a social and political object of great power, and of the assemblage of people, practices and processes that demand that it exist.

I note that several terms used in this chapter are themselves in flux and contested. These include, but are not limited to, the very definition of 'archive', particularly when juxtaposed with 'database'. Library studies scholar Marlene Manoff (2010, 385) attests to this fact in her work:

> When scholars outside library and archival science use the word 'archive' or when those outside information technology fields use the word 'database,' they almost always mean something broader and more ambiguous than experts in these fields using those same words. The disciplinary boundaries within which these terms have been contained are eroding…But archive and database have also evolved into increasingly contested terms used to theorize digital culture and new forms of collective memory.

This chapter wrestles with the kind of slippage and application of metaphor that Manoff describes in its treatment of the case of PhotoDNA, a largely automated, algorithm-reliant product that exists in a definitional boundary space and to contested ends.

AI to the Rescue?

To combat the dual problem of legal liability and brand damage from the worst kind of content, as well as content moderation worker burnout and harm from exposure (Hadley 2017), mainstream platforms have increasingly pinned their hopes on the potential of automation via algorithm to address content they wish to remove from their sites before users ever see it. The hope is that computational mechanisms based on machine learning (ML) and artificial intelligence (AI) could provide a way to fully extract the human element from the production chain, while ensuring totally routinized adherence to internal adjudication policy. Such an approach to solving platforms' content problems, however, does little to nothing to stem the uploading of user-generated content (UGC) or the impetus behind it, but instead addresses its inclusion on the platform once it has already been added. It is, in short, a largely technological solution to a complex assemblage of social problems (Gillespie 2017). This sort of technological solutionism is typically favoured by Silicon Valley, which frequently imagines AI as preferable, and a sense that there is inequity in the application of rules, norms and procedures when humans are involved and when rote and reproducible results are desired (Klonick 2018). Yet, the application of such tools is predicated on inherent abstraction and flattening of meaning, resulting in a reduction of all human processes into pastiche made up of an algorithmic, flow-chartable, if-then logic structure. It is a logic better suited to some situations than others.

Unfortunately for Silicon Valley firms subscribing to this view, ML and AI-based automation are not at the point where they can reliably fully take over moderation functions, on both technological and economic grounds. This is due in large part to the nature of much, but not all, UGC: newly-generated material that a user has created that exists nowhere else in the world, containing a complex combination of symbols, imagery and other cultural artefacts that together convey overall meaning. For content like this, it is still cheaper and more expedient to apply a human evaluation if ever that content is flagged (Crawford and Gillespie 2016).

Yet in cases where objectionable content is already known to exist and has been uploaded and removed in the past, there is a solution. It applies to the most difficult content with which commercial content moderators contend, and which creates the greatest legal liability for platforms: child sexual exploitation and abuse material. For that, PhotoDNA was created.

PhotoDNA

In 2008, Dartmouth computer scientist Hany Farid was invited by Microsoft and the US National Center for Missing and Exploited Children

(NCMEC)[1] to a meeting of technology and social media firms contending with the circulation of child sexual exploitation material – capturing the sexual abuse of children – on their platforms. Although these firms varied in their products and market positions, they all fundamentally relied upon attracting and maintain a user base via the capture and sharing of UGC, some subset of which was this disturbing, illegal material.

The intractability of the problem of removing the content from platforms was identified as existing in two registers: technological and economic. As Farid (2018, 594) described in an article on PhotoDNA's history and technology, 'throughout the day of that first meeting, I repeatedly heard that it is incredibly difficult to automatically and efficiently scrub [child sexual exploitation material] from online platforms without interfering with the business interests of the titans of tech represented in the room.' In other words, Farid was keenly aware that one way to contend with the seemingly unending problem of disturbing material being uploaded as UGC would be to slow that firehose down, give more resources over to human moderation, and change the content-based business model on platforms. Due to the economic interests invested in carrying on with the

1 The National Center for Missing and Exploited Children is a US non-governmental organization that has direct and long-standing ties to both US and international law enforcement, as well as significant partnerships both in industry and with similar groups located around the world ('National and International Collaboration', accessed 25 March 2018, http://www.missingkids.com/supportus/partners/collaboration).

status quo, it was clear to him that this route was not one that was even up for debate. Instead, the group focused its attention on a technological solution—one that for his part, Farid felt was likely out of the realm of technological possibility.

Yet one peculiar feature of a large portion of online child sexual exploitation material is that it is frequently both extant and known to law enforcement and groups such as NCMEC. Indeed, NCMEC already possessed a repository of some one million images and videos at the time of the meeting (Farid 2018, 594). Although still a complicated computer science problem to automate its identification and removal, the fact that there was a significant corpus of material against which new UGC could be compared meant that Farid was able to foresee a means of computational automation to at least deal with the recirculation of known material. Farid developed a process of ascribing hashing algorithms to images in the database of known child sexual exploitation content that could then be automatically compared to uploaded UGC on any subscribing platform or service. The powerful breakthrough for Farid's technology was that it was able to contend with successfully automating the material's identification even when the original image had been altered, edited or compressed. The end result was a product that could be deployed to automate the moderation process of this pernicious and disturbing type of UGC. As Farid (2018, 596) explained:

> After a year and a half of development and testing, photoDNA [sic] was launched in 2009 on Microsoft's SkyDrive and search engine Bing. In 2010, Facebook deployed photoDNA on their entire network. In 2011, Twitter followed suit, while Google waited until 2016 to deploy. In addition to these titans of technology, photoDNA is now in worldwide deployment. In 2016, with an NCMEC-supplied database of approximately 80,000 images, photoDNA was responsible for removing over 10,000,000 [child sexual exploitation] images, without any disputed take-downs. This database could just as easily be three orders of magnitude bigger, giving you a sense of the massive scale of the global production and distribution of [child sexual exploitation].

An Uncertain Archive of Abuse

The PhotoDNA project has done much to curb the circulation and distribution of child sexual exploitation imagery known to law enforcement for those mainstream social media platforms that use the service. It has also taken the human commercial content moderators out of the loop of needing to review these known bad images and videos, one of the most difficult aspects of the job. Nevertheless, there are drawbacks, some of which Farid alludes to in his article on the project. As he notes, very little can be done to automate removal of material that has been newly produced or is otherwise unknown to the PhotoDNA digital archive. For this removal, commercial content moderators are still the front line;

indeed, every moderator I have ever spoken to in the course of my nine years researching this labour has indicated to me that he or she has witnessed and dealt with this material.

Per Farid, this is a problem unlikely to abate, and neither content moderators nor AI are likely to be able to change aspects of human nature that compel people to abuse others and then trade in depictions of that abuse. But it also remains to be fully understood to what extent the platforms themselves have provoked an impetus for such material to be generated and circulated, given that indiscriminate and endless UGC upload, circulation and consumption is at the core of their revenue generation. In short, there is a subset of the world's population that engages in child sexual exploitation material as producers and consumers, and they have found an expedient and powerful mechanism to circulate that material via social media and other UGC-reliant platforms (such as file-sharing tools).

Paradoxically, the solution that platforms and computer scientists have developed is not to remove all traces of this material, which can frequently be used as evidence in criminal prosecutions, and the mere possession and circulation of which typically constitutes a crime. Instead, a fundamental part of the process of dealing with the removal of child sexual exploitation content on a mainstream social media site is, bizarrely, to archive it: it is subsumed into the PhotoDNA digital database to be catalogued, hashed and used for matching purposes against the

new uploads that never cease. For these reasons, digital records of someone's victimization will exist in perpetuity. It is not clear to what extent the victims themselves are aware of this fact. There is eusconflation, too, of the business needs of the UGC hosts with a legal and moral responsibility to intervene upon this material. As Microsoft's own PhotoDNA landing page proclaims, 'Help stop the spread of child exploitation images and protect your business,'[2] suggesting a practical, if not moral, equivalency between the two.

There is a deep uncertainty in the resulting archive produced by technologies such as PhotoDNA: it is fundamentally unknown, unknowable, inaccessible, and exists expressly not to be seen or even apprehended, and yet it reflects the tendency—fomented or at least facilitated by digital technologies—to collect, sift through, categorize, catalogue and possess (Bowker and Star 1999). It has other peculiar features, too: it is an archive of material meant to never be seen but instead to be collected so as to remove it from circulation; this archive as object, as well as its particular functionality, is invisible to the public. Records within it are not categorized and catalogued for user-facing findability or usability, but rather for the purposes of rescinding even more like material from view.

2 'PhotoDNA Cloud Service', accessed 31 December 2017, https://www.microsoft.com/en-us/photodna.

In that sense, it grows by subtraction: removal of material from user-facing accessibility means growth of the archive. It uses hashing to automate its growth, but new material must also be added as well. The systematic, always-on nature of the removal process is hidden. The material, *in toto*, constitutes an archive of absence that is predicated on a larger logic of opacity upon which social media UGC is solicited, monetized and circulated, and is the undergirding logic of the economics of mainstream social media platforms (Roberts 2018). Ultimately, PhotoDNA exists as something like a black hole—we can conceptualize its borders or even feel a certain gravitational pull or flow towards it, but can never and must never delve inside. It is a vortex at once on the periphery of social media's operational structure and at the same time central to it.

Media scholar Abigail de Kosnik (2016) has proposed the notion of 'rogue archives' in her work; in that context, she refers to the collective output of fandom communities who create material (remixed or newly generated) outside of the auspices of officially sanctioned institutional archives, and often outside of professional communities of archival practice. To what extent can PhotoDNA be considered a rogue archive of its own? Perhaps, beyond even rogue, it is the un-archive—that archive that is not one, at least not in any traditional sense. And yet what else but an archive can this collection of related, collected, sorted and catalogued artefacts be called? These are questions that must be addressed in order to make any sense of the social meaning of PhotoDNA and the power relations it implies.

Conclusion: Reading Technologies

The field of critical archival studies has offered numerous cases of challenging and difficult archives, particularly those that address or serve as repositories for human rights abuses. In this light, archives can serve to powerfully bear witness to abuses of power, but at the same time occupy the complex role of rendering those abuses painfully and repeatedly visible. As Caswell (2013, 605) asserts: ‘Contrary to positivist conceptions, records aren’t neutral by-products of activity; they are discursive agents through which power is made manifest. Records both produce and are produced by violent acts.’

Where, then, does the power lie in PhotoDNA, and in the service of whom? And, as Caswell and others have argued, if various manifestations of archives can exist to rework a power imbalance (e.g., community archives), particularly in the case of human rights violations, to what extent, if any, does PhotoDNA do that? Caswell (2014b), for example, has put forth the notion of a ‘survivor-centred approach to records’ in cases of documenting human rights violations, and yet, in the case of PhotoDNA, the survivors themselves seem to be largely absent from consideration altogether.

Perhaps this is due to the fact that, at present, PhotoDNA exists as an unpleasant and frightening outcome of UGC as economic model, and yet, like commercial content moderation work itself, is often thought of as

aberration when publicly discussed at all. The fact that this outcome may be possible to avoid is never taken under advisement; the UGC-reliant social media economic model that encourages or at least facilitates the circulation of such imagery and material is never seriously questioned by those firms who require PhotoDNA. And because what is captured and subsumed into PhotoDNA—and PhotoDNA itself—is largely imperceptible to the average user, it becomes difficult for anyone to seriously think about the social role of the PhotoDNA archive, or that of UGC-based social media platforms in general, with full information.

But such informed readings, to return to Ketelaar, will become even more key as more and more material finds its way, through manual or automated means, into one-directional repositories like PhotoDNA, such as in the case of ‘terrorism’ content (Thakor 2016). In the spirit of scholars who have taken up the critical question of the social role and power of the archive (Caswell 2014a), blockchain (Golumbia 2016), algorithms (Bucher 2017), supply chains (Posner 2018) and search (Noble 2018), this essay issues an invitation to collectively unpack the social role of UGC removal in a holistic sense, from the humans who undertake the process by hand to the automated tools that may one day largely supersede them, and to the impact on the resulting social media ecosystem that is irrevocably shaped by these presences and absences.

Bibliography

Bowker, G., and S.L. Star. 1999. *Sorting Things Out: Classification and Its Consequences*. Cambridge, MA: MIT Press.

Bucher, T. 2017. ‘The Algorithmic Imaginary: Exploring the Ordinary Affects of Facebook Algorithms’. *Information, Communication & Society* 20 (1): 30–44. doi:10.1080/1369118X.2016.1154086.

Caswell, M. 2013. ‘Not Just Between Us: A Riposte to Mark Green’. *American Archivist* 76 (2): 605–8.

———. 2014a. *Archiving the Unspeakable: Silence, Memory, and the Photographic Record in Cambodia*. Madison: University of Wisconsin Press.

———. 2014b. ‘Toward a Survivor-Centered Approach to Records Documenting Human Rights Abuse: Lessons from Community Archives’. *Archival Science* 14 (3–4): 307–22. doi:10.1007/s10502-014-9220-6.

Crawford, K., and T. Gillespie. 2016. ‘What Is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint’. *New Media & Society* 18 (3): 410–28.

Farid, H. 2018. ‘Reining in Online Abuses’. *Technology & Innovation* 19 (3): 593–99. doi:10.21300/19.3.2018.593.

Gillespie, T. 2017. ‘Governance of and by Platforms’. In *Sage Handbook of Social Media*, edited by J. Burgess, A. Marwick and T. Poell, 30. London: Sage.

Golumbia, D. 2016. *The Politics of Bitcoin: Software as Right-Wing Extremism*. Minneapolis, MN: University of Minnesota Press.

Hadley, G. 2017. ‘Forced to Watch Child Porn for Their Job, Microsoft Employees Developed PTSD, They Say’. *McClatchy DC*, 11 January. http://www.mcclatchydc.com/news/nation-world/national/article125953194.html.

Ketelaar, E. 2001. ‘Tacit Narratives: The Meanings of Archives’. *Archival Science*, 1 (2): 131–41. doi:10.1007/BF02435644.

Klonick, K. 2018. ‘The New Governors: The People, Rules, and Processes Governing Online Speech’. *Harvard Law Review*, (131), 1598–1670.

Kosnik, A.D. 2016. *Rogue Archives: Digital Cultural Memory and Media Fandom*. Cambridge, MA: MIT Press.

Manoff, M. 2010. ‘Archive and Database as Metaphor: Theorizing the Historical Record’. *Portal: Libraries and the Academy*, 10 (4): 385–98. doi:10.1353/pla.2010.0005.

Myers West, S. 2018. ‘Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms’. *New Media & Society*. doi:10.1177/1461444818773059.

Noble, S.U. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Posner, M. 2018. ‘See No Evil’. *Logic* 1 (4): 215-229.

Punzalan, R.L., and M. Caswell. 2015. ‘Critical Directions for Archival Approaches to Social Justice’. *Library Quarterly* 86 (1): 25–42. doi:10.1086/684145.

Roberts, S.T. 2018. ‘Digital Detritus: Bodies, “Error” and the Logic of Opacity in Social Media Content Moderation’. *First Monday*. doi: http://dx.doi.org/10.5210/fm.v23i3.8283.

Sutherland, T. 2017. ‘Archival Amnesty: In Search of Black American Transitional and Restorative Justice’. *Journal of Critical Library and Information Studies* 1 (2). doi:10.24242/jclis.v1i2.42.

Thakor, M.N. 2016. ‘Algorithmic Detectives Against Child Trafficking: Data, Entrapment, and the New Global Policing Network’. PhD diss., Massachusetts Institute of Technology. http://dspace.mit.edu/handle/1721.1/107039.

Wood, S., K. Carbone, M. Cifor, A. Gilliland and R. Punzalan. 2014. ‘Mobilizing Records: Re-Framing Archival Description to Support Human Rights’. *Archival Science* 14 (3–4): 397–419. doi:10.1007/s10502-014-9233-1.